基于深度学习的关节点行为识别综述

刘 云* 薛盼盼 李 辉 王传旭
(青岛科技大学信息科学技术学院 青岛 266061)

摘 要：关节点行为识别由于其不易受外观影响、能更好地避免噪声影响等优点备受国内外学者的关注，但是目前该领域的系统归纳综述较少。该文综述了基于深度学习的关节点行为识别方法，按照网络主体的不同将其划分为卷积神经网络(CNN)、循环神经网络(RNN)、图卷积网络和混合网络。卷积神经网络、循环神经网络、图卷积网络分别擅长处理的关节点数据表示方式是伪图像、向量序列、拓扑图。归纳总结了目前国内外常用的关节点行为识别数据集，探讨了关节点行为识别所面临的挑战以及未来研究方向，高精度前提下快速行为识别和实用化仍然需要继续推进。

关键词：深度学习；关节点行为识别；卷积神经网络；循环神经网络；图卷积
中图分类号：TN911.73; TP391 **文献标识码**：A **文章编号**：1009-5896(2021)06-1789-14
DOI: 10.11999/JEIT200267

A Review of Action Recognition Using Joints Based on Deep Learning

LIU Yun XUE Panpan LI Hui WANG Chuanxu
(*College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao* 266061, *China*)

Abstract: Action recognition using joints has attracted the attention of scholars at home and abroad because it is not easily affected by appearance and can better avoid the impact of noise. However, there are few systematic reviews in this field. In this paper, the methods of action recognition using joints based on deep learning in recent years are summarized. According to the different subjects of the network, it is divided into Convolutional Neural Network(CNN), Recurrent Neural Network(RNN), graph convolution network and hybrid network. The representation of joint point data that convolution neural network, recurrent neural network and graph convolution network are good at is pseudo image, vector sequence and topological graph. This paper summarizes the current data sets of action recognition using joints at home and abroad, and discusses the challenges and future research directions of behavior recognition using joints. Under the premise of high precision, rapid action recognition and practicality still need to be further promoted.

Key words: Deep learning; Action recognition using joints; Convolution Neural Network(CNN); Recurrent Neural Network(RNN); Graph convolution

1 引言

人类行为识别是计算机视觉的一个重要分支，在很多方面都有广泛应用，例如智能监控、人机交互、视频检索和运动分析[1]。目前，已有一些学者对行为识别进行了综述，比如朱煜等人[2]、罗会兰等人[3]、张会珍等人[4]、Zhu等人[5]，这些综述文章无论是侧重于传统行为识别方法还是侧重于深度学习行为识别方法，所利用的信息多是RGB(Red (红色)、Green(绿色)、Blue(蓝色))数据和深度数据，没有专门针对关节点信息行为识别进行系统的归纳总结。近年来，关节点数据的获取随着低成本设备的发展更加容易，例如Microsoft Kinect[6]。随着深度学习的发展，利用关节点数据进行行为识别的研究取得了丰硕成果，但目前在该领域的系统归纳较少。与RGB数据和深度数据相比，关节点本身是人体的高级特征，不易受外观影响，同时能够更好地避免背景遮挡、光照变化以及视角变化产生的噪声影响，同时在计算和存储方面也是有效的[7]。利用关节点数据进行行为识别从发展历程上主要分为两大类：基于手工特征的方法和基于深度学习的方法。传统的利用关节数据进行行为识别是基于手工特征[8–10]。

收稿日期：2020-04-14；改回日期：2020-12-30；网络出版：2021-01-11
*通信作者： 刘云 lyun-1027@163.com
基金项目：国家自然科学基金(61702295, 61472196)
Foundation Items: The National Natural Science Foundation of China (61702295, 61472196)

关节点数据通常表示为一系列点的坐标向量，在不同的深度学习网络和算法中，关节点数据一般表示为伪图像、向量序列和拓扑图，不同的深度学习主干网络架构适合处理的数据表示方式也不同。通常来说，基于深度学习算法的改进主要是针对3个方面：数据处理方式、网络架构和数据融合方式。数据处理方式主要表现为是否进行数据预处理和数据降噪的方法，不同技术之间的数据融合方式也较为相似，对研究工作区分较大的是网络架构，因此本文也将根据主干网络架构的不同对关节点行为识别方法进行归纳总结。

2 基于深度学习的关节点行为识别

在深度学习背景下，关节点行为识别是针对已剪辑好的包含关节点位置数据的视频片段进行的特征提取和识别。常见处理关节点数据的深度学习方法有卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)、图卷积网络，对应的关节点数据的表示方式为伪图像、向量序列和拓扑图。本节按照主干网络将基于深度学习的关节点行为识别方法分为基于卷积神经网络的关节点行为识别、基于循环神经网络的关节点行为识别、基于图卷积网络的关节点行为识别和基于混合网络的关节点行为识别。图1为基于深度学习的关节点行为识别流程图。首先原始的关节点数据输入网络，其中横轴方向表示关节点的编号，纵轴方向的(x,y,z)表示关节点的3维坐标，竖轴方向表示时间帧，然后将其馈送到不同的网络模型中进行行为特征的提取，最终得到行为识别结果。

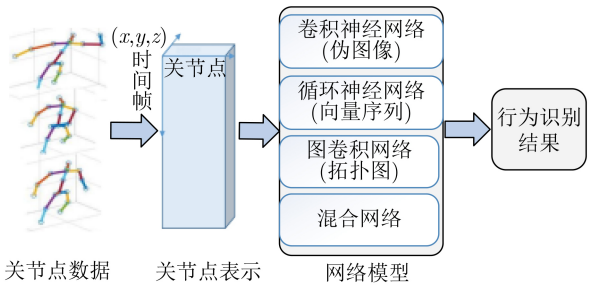

图 1 基于深度学习的关节点行为识别流程图

2.1 基于卷积神经网络的关节点行为识别

CNN提供了一种有效的网络架构，可以在大型数据集中提取人体行为特征，这些特征可通过从数据中学习到的局部卷积滤波器或内核来识别。基于CNN的方法分别将时间帧和骨架关节的位置坐标编码为行和列，然后将数据馈送到CNN中进行行为识别，类似于图像分类。图2为基于卷积神经网络的关节点行为识别流程图。首先将原始的关节点数据输入到行为识别网络中，一般为了方便使用基于CNN的网络做特征提取会将关节点数据进行转置映射到图像中，其中行表示不同的关节，列表示不同的帧，(x,y,z)的3D坐标值被视为图像的3个通道，然后进行卷积操作。卷积展开的作用是将多维的数据1维化，该环节是卷积操作和全连接之间的常用过渡方式。全连接是在整个卷积神经网络中起到“分类器”的作用，也就是将学到的特征空间表示映射到样本标记空间。最后经过这一系列的操作就能够得到行为识别的结果。

Li等人[11]所提出的平移尺度不变图像映射和多尺度深度CNN网络是一个比较有代表性的基于卷积神经网络的关节点行为识别方法。一般来说将3D骨架视频映射到图像中较为直接的方法是将关节点的坐标分量(x,y,z)表示为每个像素对应的分量(R,G,B)，动作图像的每一行被定义为$\boldsymbol{R}_i=[x_{i1},x_{i2},\cdots,x_{iN}]$，$\boldsymbol{G}_i=[y_{i1},y_{i2},\cdots,y_{iN}]$，$\boldsymbol{B}_i=[z_{i1},z_{i2},\cdots,z_{iN}]$，其中$i$表示关节点索引，$N$表示时间序列中的帧数。通过这种方式将所有关节点连接在一起得到原始骨架的动作图像表示。由于3D骨架和图像的坐标差异，作者提出了一种简单而有效的平移尺度不变的映射方法，如式(1)所示

$$p^{jk}=\text{floor}\left(255\times\frac{c^{jk}-c^{jk}_{\min}}{\max\limits_{k}(c^{jk}_{\max}-c^{jk}_{\min})}\right) \tag{1}$$

其中，$c^{jk}_{\max}$和$c^{jk}_{\min}$是第j帧视频序列中的(x,y,z)的第k通道的最大和最小的坐标值，x, y, z分别对应的

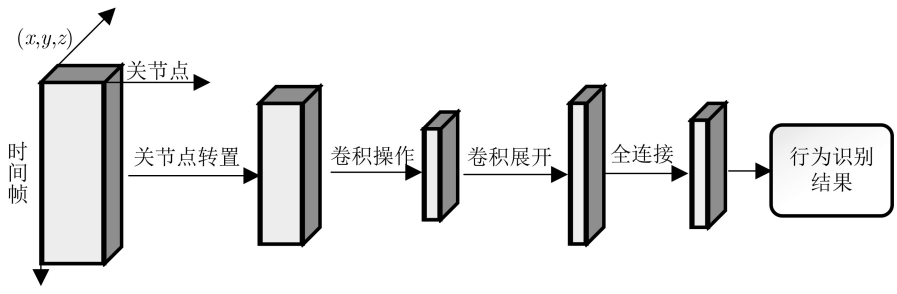

图 2 基于卷积神经网络的关节点行为识别流程图

是第1, 2, 3通道。通过这种平移尺度不变的映射方法不仅能将关节点数据的视频信息标准化到0～255的范围，还能保证在映射到彩色图像的同时实现平移和比例不变性以及数据集独立，该方法在很多基于卷积神经网络的关节点行为识别方法中应用。将骨架的关节点信息映射到图像中再采用多尺度深卷积神经网络结构和微调策略来提高性能，该架构能够在预训练的CNN上进行，可以较大程度地提高训练效率和准确性。

2017年Kim等人[12]提出一种残差时间卷积网络用于关节点行为识别，该网络框架是一种明确学习易于解释的3D人类行为识别的时空表示方法。残差时间卷积是在CNN的基础上设计的，网络由1维卷积的堆叠单元构成，并且能够在时间和空间上分配不同程度的注意力，但是该方法的识别精度一般。同年Li等人[13]采用双流CNN架构组合人体关节的位置和速度信息，同时引入了一种新的骨架变换器模块，实现了重要骨架节点的自动重新排列和选择，该方法较高的识别准确率证明了CNN模拟时间模式的能力。Liu等人[14]提出视图不变方法，不仅消除视图变化的影响还能保留原始关节数据中的运动特征，同时提出一种增强的骨架可视化方法用于视图不变的人体行为识别。Ke等人[15]于2017年最先将迁移学习应用于关节点行为识别中。同年Ke等人[16]又进一步将原始关节点数据转换为3个灰度图像片段，灰度图像是使用关节与参考关节之间的相对位置生成的，这与Li等人[13]的转换方法类似，Ke等人[16]所提出的方法首先将每个骨架序列转换成3个片段，每个片段由几帧组成，用于使用深度CNN进行空间时间特征学习，识别准确率提高了约4%。由于先前的研究并未完全利用人体行为中视频片段之间的时间关系，Le等人[17]在2018年提出了一种新的框架，该框架首先将骨架序列分割为不同的时间段，然后利用从细到粗的CNN架构同时提取关节点序列的时间和空间特征。该网络架构较浅，能够一定程度上避免数据量不足的问题，从表1可以看出，在SBU这种不是特别大的数据集上识别精度很好，达到了99.1%。Li等人[18]提出层次共现网络，首先将每个关节点进行单独的编码，用CNN独立地学习每个关节点的点水平特征，然后将每个关节都视为CNN的通道来学习层次共现特征，其行为识别准确率超越了大多数基于卷积神经网络的关节点行为识别方法。刘庭煜等人[19]针对生产车间工作人员行为识别与智能监控问题提出一种基于关节点数据的生产行为识别方法，首先将预处理好的人体关节点数据合并成人体行为的时空特

表 1　主干网络为卷积神经网络的关节点行为识别及代表性工作

年份	技术特点	模型优劣分析	实验结果(%)		
			NTU RGB+D	SBU	JHMDB
2017	平移尺度不变图像映射和多尺度深度CNN[11]	可以在预训练的CNN网络上进行	CS:85.0 CV:96.3	–	–
2017	残差时间卷积[12]	模型易于解释，但准确率一般	CS:74.3 CV:83.1	–	–
2017	引入骨架变换的双流CNN架构[13]	证明了CNN具有时间模拟能力	CS:83.2 CV:89.3	–	–
2017	多流卷积神经网络[14]	消除视图变化的影响且保留原始关节数据中的运动特征	CS:80.0 CV:87.2	–	–
2017	卷积神经网络[15]	将迁移学习应用于关节点行为识别，提高了训练效率	CS:75.9 CV:81.2	–	–
2017	卷积神经网络+多任务学习[16]	训练效率低	CS:79.6 CV:84.8	Acc:93.6	–
2018	从细到粗的卷积神经网络[17]	网络架构较浅，能避免数量不足容易过拟合的问题	CS:79.6 CV:84.6	Acc:99.1	–
2018	分层共现的卷积神经网络[18]	能利用不同关节之间的相关性	CS:86.5 CV:91.1	Acc:98.6	–
2019	双流的卷积神经(RGB信息和关节点信息结合)[20]	训练时间短	CS:80.09	Acc:92.55	–
2019	卷积神经网络(多姿势模态)[21]	网络框架简洁，准确率一般	–	–	Acc:69.5
2019	卷积神经网络(树结构和参考关节的图像表示方法)[22]	训练效率不高	–	–	Acc:69.5
2019	卷积神经网络(重新编码骨架关节的时间动态)[23]	能够有效过滤数据中的噪声	CS:76.5 CV:84.7	–	–
2019	卷积神经网络(轻量级)[24]	速度快，准确率低	CS:67.7 CV:66.9	–	Acc:78.0

征RGB图像，然后送入3维深度卷积神经网络中，该方法具有较高实用价值，并且在数据集MSR Action3D上的准确率可以达到84.27%。针对复杂的交互动作识别准确率不够高的问题，姬晓飞等人[20]提出一种基于RGB和关节点数据双流信息融合的卷积神经网络，其中RGB视频信息在送入卷积神经网络之前进行关键帧的提取缩短了训练时间，双流信息的融合提高了识别准确率。Yan等人[21]提出基于姿态的行为识别网络，该网络是一个简洁3维CNN框架，由空间姿态CNN、时序姿态CNN和动作CNN 3个语义模块组成，可以作为另一个语义流与RGB流和光学流互补，该网络框架较为简洁，但是准确率一般，在JHMDB数据集上的准确率仅为69.5%。Caetano等人[22,23]、Li等人[24]从设计新的骨架表示图像入手，其中Caetano等人[22]提出一种基于树结构和参考关节的3维行为识别的骨架图像表示方法，在JHMDB数据集上的识别准确率与Yan等人[21]所提出方法相同。Caetano等人[23]又引入了一种新的方法通过计算骨架关节的运动幅度和方向值来编码时间动态，使用不同的时间尺度来计算关节的运动值能够有效过滤噪声运动值。Li等人[24]是用集合代数的方式对骨架关节信息进行重新编码。Yang等人[25]提出了一个轻量级的网络框架，该网络由多个卷积神经网络组合而成，大大提高了速度，但是识别精度和其他方法相比较低。主干网络为卷积神经网络的关节点行为识别及代表性工作如表1所示。

2.2 基于循环神经网络的关节点行为识别

循环神经网络(RNN)可以处理长度可变的序列数据，长短期记忆模型(Long Short Term Memory, LSTM)是一种变种的RNN，由于其细胞状态能够决定哪些时间状态应该被留下哪些应该被遗忘，所以在处理关节点视频这种时序数据时有更大优势，从而被较多地应用到关节点行为识别中，图3为基于循环神经网络的行为识别流程图。首先将关节点数据表示为向量序列，每一个向量序列包含一个时间帧上的所有关节点的位置信息；然后将向量序列送入以循环神经网络为主干的行为识别网络中；最后得到行为识别的结果。

Shahroudy等人[26]在2016年提出了NTU RGB+D数据集，同时提出了一种新的递归神经网络来模拟每个身体部位特征的长期时间相关性进行关节点数据的行为识别，可以更有效并且直观地保持每个身体部位的上下文信息，但是识别准确率不高，在NTU RGB+D数据集上跨表演者模式(Cross Sbuject, CS)的准确率是62.9%，跨视角模式(Cross View, CV)的准确率是70.3%。该文献为之后利用NTU RGB+D数据集进行行为识别研究的方法提供了对比的基准。Liu等人[27]提出一种基于信任门的长短期记忆模型(SpatioTemporal-Long Short Term Memory, ST-LSTM)，信任门模块能够降低关节点数据的噪声，提高行为识别的准确率。Liu等人[28]又在ST-LSTM的基础上做了进一步的改进，在LSTM中加入一种新颖的多模式特征融合策略，使在多个标准数据集上的准确率(比如NTU RGB+D和UK-Kinect)都有较大提升，其中在NTU RGB+D数据集上的准确率提高了约3%。2017年Liu等人[29]提出全局上下文感知长短期记忆模型框架(Global Context-aware Attention Long Short Term Memory networks, GCA-LSTM)，该框架主要由两层LSTM构成，第1层生成全局的背景信息，第2层加入注意力机制，更好地聚焦每一帧的关键关节点从而提高行为识别准确率。同年Liu等人[30]又在GCA-LSTM的基础上进行了扩展，加入粗粒度和细粒度的注意力机制，识别准确率在NTU RGB+D数据集上约提高了3%，在UK-Kinect数据

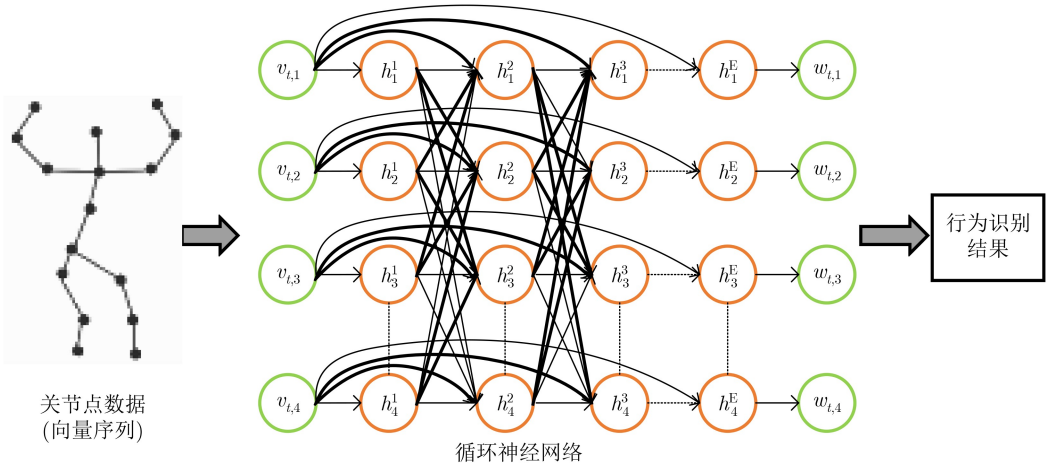

图 3 基于循环神经网络的行为识别流程图

集上提高了约1%。Zheng等人[31]提出了一种双流注意力循环LSTM网络，如图4所示。循环关系网络学习单个骨架中的空间特征，多层LSTM学习骨架序列中的时间特征。该双流的网络中，一个网络输入的是原始关节点数据，另一个网络输入的是成对关节之间的连线数据，关节点数据强调绝对位置，连线数据强调相对位置。在每个流中，首先增加每个关节点或关节连线数据的维数，然后发送给RNN用于提取单个骨架中的空间特征，同时生成一个可学习的掩码将更多注意力集中在骨架的潜在区分部分，再使用多层LSTM学习骨架序列的时间特征，最后以加权平均运算作为融合策略，以合并来自两个流的预测。该网络能更加有效地利用丰富的结构或关节信息，准确率较高。Li等人[32]提出了一个独立递归神经网络(Independently Recurrent Neural Network, IndRNN)，不同层之间的神经元之间跨层连接，同一层中的神经元彼此独立，能更好地在网络较深的情况下防止梯度爆炸和梯度消失。王佳铖等人[33]针对车间作业行为识别问题提出了基于工件注意力的车间行为在线识别模型，不仅通过将人的关节点信息输入以门控循环单元为基础的模型对行为动作进行分类，还同时将工件的语义特征作为注意力融入进去，该方法有利于提高车间数字化管理能力，最终在自建数据集上准确率为88.5%，但是在标准数据集IXMAS上准确率仅为29.8%，这说明该方法适用性较差。主干网络为循环神经网络的关节点行为识别及代表性工作如表2所示。

2.3　基于图卷积网络的关节点行为识别

人体骨架关节本身是一种拓扑图，卷积神经网络无法直接处理这种非欧几里得结构的数据，因为

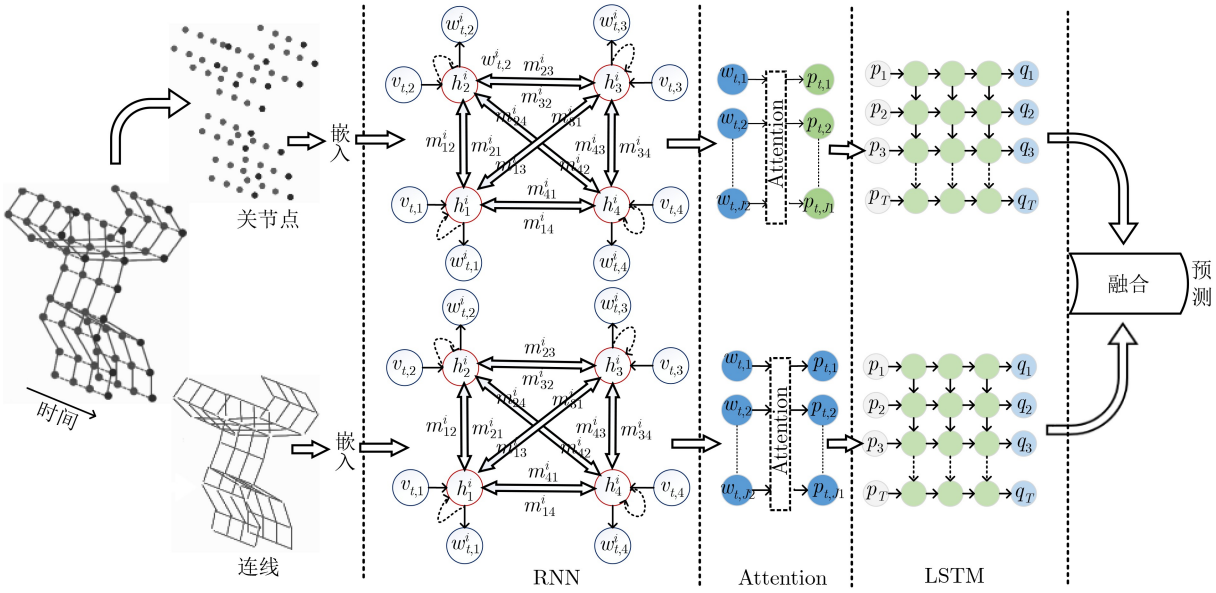

图 4 双流长短期记忆模型框架[31]

表 2　主干网络为循环神经网络的关节点行为识别及代表性工作

年份	技术特点	模型优劣分析	实验结果(%)		
			NTU RGB+D	UK-Kinect	SYSU 3D
2016	长短期记忆模型(将身体分为5个部分)[26]	能有效且直观地保持上下文信息，但是识别准确率不高	CS: 62.9 CV:70.3	–	–
2016	基于信任门的长短期记忆模型[27]	能够降低关节点数据的噪声	CS:69.2 CV:77.7	Acc:97.0	–
2017	基于信任门的长短期记忆模型(加入多模式特征融合策略)[28]	提高了识别准确率，降低了训练效率	CS:73.2 CV:80.6	Acc:98.0	Acc:76.5
2017	全局上下文感知长短期记忆模型(注意力机制)[29]	能够更好地聚焦每一帧中的关键关节点	CS:74.4 CV:82.8	Acc:98.5	–
2017	全局上下文感知长短期记忆模型(双流+注意力机制)[30]	提高了识别准确率，降低了训练效率	CS:77.1 CV:85.1	Acc:99.0	Acc:79.1
2019	双流长短期记忆模型(注意力机制)[31]	更充分地利用关节信息，提高识别准确率	CS:81.8 CV:89.6	–	–
2018	独立递归神经网络[32]	能更好地在网络较深的情况下避免梯度爆炸和梯度消失	CS:81.8 CV:88.0	–	–

拓扑图中每个点的相邻顶点数目可能不同，难以用一个同样大小的卷积核进行卷积计算，而图卷积神经网络能够直接处理这种拓扑图。图5为基于图卷积网络的行为识别流程图。首先将关节点数据表示为拓扑图，在空间域上顶点由空间边缘线连接，在时域上相邻帧之间对应关节由时间边缘线连接，每个关节点的属性特征是空间坐标向量；然后将拓扑图输入以图卷积网络为主干的行为识别网络中，最终得到行为识别的结果。

Yan等人[34]在2018年提出了基于骨架的行为识别时空图卷积网络，这是第1篇用图卷积网络进行关节点行为识别的文章，同时也是非常有代表性的一个方法，主要介绍了如何利用人体关节点构造图以及在所构造的图上进行图卷积操作。Yan等人[34]构建图的规则是在空间域上根据人体关节的自然连接构造图，在时间域上是在连续的帧中添加相应关节之间的时间边。对于单个人的单帧骨架可以表示为$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，其中$\boldsymbol{V}$是关节点的集合，$\boldsymbol{E}$是边的集合。针对一系列时间帧的骨架可以表示为$\boldsymbol{F}=\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_T$，其中$T$表示视频总的帧数。顶点$v_i$在空间域进行图卷积如式(2)所示

$$f_{\text{out}}(v_i)=\sum_{v_j\in D_i}(1/Z_{ij})f_{\text{in}}(v_j)*w(l_i(v_j)) \tag{2}$$

其中，f表示特征图，v表示图的顶点，D_i表示第i个顶点v_i的采样区域，即v_i的邻居顶点v_j的集合，w是加权函数，l_i是权重的映射函数。卷积的权重向量的个数由映射策略中把D_i划分的子集的个数所决定，该文章中所采用的映射策略是按照空间结构把D_i划分为3个子集：根节点(顶点v_i)、向心点、离心点。其中向心点是到骨架重心的距离比根节点到骨架重心距离更短的关节点，离心点是到骨架重心的距离比根节点到骨架重心更长的关节点。Z_{ij}表示第i个顶点的第j个子集的基数，$1/Z_{ij}$可以平衡每个子集的贡献。

Yan等人[34]使用图卷积进行关节点行为识别能够形成骨架关节的层次表示得到较好的识别结果，但由于感受野较小，难以学习无物理联系的关节之间的关系。Shi等人[35]、Li等人[36]都在试图克服这些问题，Shi等人[35]提出的双流自适应图卷积网络，骨架关节的拓扑图可以用BP算法自适应地学习，增加图形构建模型的灵活性。该双流框架不仅利用骨架数据的1阶信息(关节点信息)，还利用骨架的2阶信息(骨骼的长度和方向)，在NTU RGB+D数据集上准确率较Yan等人[34]的方法提高了约7%。Li等人[36]提出了一种编码器-解码器的方法来捕获隐含的关节相关性以及使用邻接矩阵的高阶多项式获取关节之间的物理结构链接。Gao等人[37]将图形回归用于基于骨架的行为识别，对于图卷积而言，图形的表示很重要，图形回归的方法能够优化时空帧的基础图形，充分利用人体关节之间空间上物理和非物理的依赖关系以及连续帧上的时间连通性。Li等人[38]提出一种时空图卷积方法，能够将自回归滑动平均序列学习能力与局部卷积滤波器结合。对于每个帧构造无向图，其中仅按照人体关节的自然连接构造图，无时间连通性，在NTU RGB+D上的识别准确率CS和CV分别为74.9%和86.3%， 与其他方法相比准确率较低。Tang等人[39]提出深度渐进强化学习方法，该方法可以提取关键帧，然后用图卷积网络进行行为识别，行为识别的准确率一般，但是提高了训练效率。在实际应用中经常遇到关节点信息缺失的问题，大多数的基于关节点行为识别的模型都是针对完整的骨架数据，但是真实场景中可能会出现部分关节点信息缺失的情况，Song等人[40]提出针对不完整骨架的行为识别的激活图卷积网络，以提高图卷积网络在关节点行为识别中的鲁棒性。Peng等人[41]提出将神经体系结构搜索用于构建图卷积网络，该搜索策略中将交叉熵演化策略与重要性混合方法相结合，提高了采样效率和存储效率。Wu等人[42]提出将空间残差层和密集连接块增强引入时空图卷积网络，这种方法能够提高时空信息的处理效率，并且也容易与主流时空图卷积方法结合。Shi等人[43]在双流自适应图卷积网络[35]的基础上进行改进，将骨架数据表示为基于自然人体关节和骨骼之间运动依赖的有向无环图，准确率提升了约1%。Li等人[44]提出了一种新颖的共生图卷积网络，该网络不仅包含行为识别的

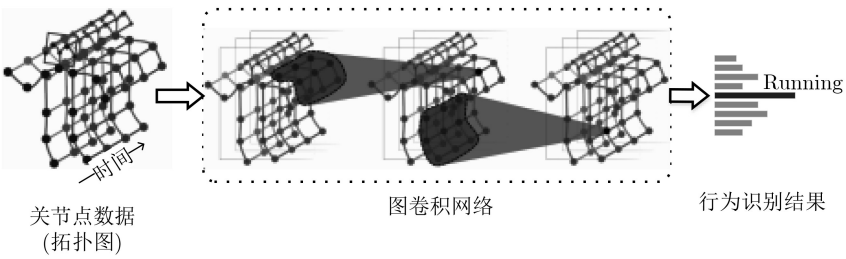

图 5 基于图卷积网络的行为识别流程图

功能模块，还包含动作预测模块，两个模块相互促进，显著提高了行为识别和动作预测的准确率，在NTU RGB+D数据集上CS和CV的准确率均超过90%。Yang等人[45]提出一个带有时间和通道注意力机制的伪图卷积网络，通过这种方式不仅能提取关键帧，还能筛选出包含更多特征的输入帧。行为识别性能优于大多数方法，但仍存在问题，因为帧数远远大于通道数，可能会导致省略一些关键信息。图卷积网络虽然能提高识别的准确率，但计算较复杂，计算速度也较慢，Wu等人[46]、Chen等人[47]更关注提高图卷积网络的速度，其中Wu等人[46]所提到的方法比Chen等人[47]所提到的方法产生高达两个数量级的加速。主干网络为图卷积网络的关节点行为识别及代表性工作如表3所示。

2.4 基于混合网络的关节点行为识别

与以上3种主干网络架构相比，基于混合网络的关节点行为识别的研究充分利用了卷积神经网络和图卷积网络在空间域上特征提取的能力以及循环神经网络在时序分类的优势，能够得到较好的行为识别结果。图6为基于混合网络的关节点行为识别流程图。首先将原始的关节点数据根据不同的混合网络的需要进行相应的关节点表示；然后将其馈送进混合网络中，混合网络的主干网络一般会包含卷积神经网络、基于循环神经网络、图卷积网络中的两个或更多；最终将提取到的行为特征进行行为分类得到行为识别结果。

Zhang等人[48]提出了一种视图自适应方案，根据该方案设计了两个视图自适应神经网络，分别基于LSTM和CNN，视图自适应子网会在识别期间自动确定最佳的虚拟观察视点。视图自适应神经网络由两大部分组成，一个是由视图自适应子网和主LSTM组成的视图自适应循环网络，将新的视点下的关节点表示送入主LSTM网络确定行为识别，如图7所示；还有一个是由视图自适应子网和主CNN组成的视图自适应卷积网络，将新的观察视点下的关节点表示送入主CNN中确定行为类别。分阶段训练完之后，再将两部分网络的分类分数融合预测。该方法不仅减弱了视角不同对行为识别结果的影响，同时利用了CNN擅长提取空间域特征和循环神经网络擅长提取时间域行为特征的优势，得到了较好的行为识别结果。Hu等人[49]不仅考虑时间域和空间域行为特征的提取，还提出了一种残差频率注意力方法，主要用来学习频率模式，该文献所提出的网络框架可以看作CNN的变体和图杂交方法结合，取得了较高的行为识别准确率。Si等人[50,51]、Gao等人[52]都是采用图卷积和LSTM相结合的方式进行关节点的行为识别研究，图卷积更加擅长空间域的特征提取，LSTM更加擅长时间域的特征提取。Si等人[51]所提出的注意力增强图卷积LSTM网络(Attention enhanced Graph Convolutional Long Short Term Memory network, AGC-LSTM)，不仅可以提取空间域和时间域的行为特征，还通

表 3　主干网络为图卷积网络的关节点行为识别及代表性工作

年份	技术特点	模型优劣分析	实验结果(%)		
			NTU RGB+D	Kinects	Florence 3D
2018	时空图卷积网络[34]	难以学习无物理联系关节之间的关系	CS:81.5 CV:88.3	Top1:30.7 Top5:52.8	–
2018	双流自适应图卷积[35]	充分利用骨架的2阶信息 (骨骼的长度的方向)	CS:88.5 CV:95.1	Top1:36.1 Top5:58.7	–
2019	图卷积(编解码)[36]	模型复杂度高	CS:86.8 CV:94.2	Top1:34.8 Top5:56.5	–
2018	时空图卷积网络(图回归)[37]	充分利用关节之间的物理和非物理的 依赖关系以及连续帧上的时间连通性	CS:87.5 CV:94.3	–	Acc:98.4
2018	时空图卷积网络[38]	缺乏时间连通性	CS:74.9 CV:86.3	–	Acc:99.1
2018	关键帧提取+图卷积网络[39]	关键帧的提取能够提高训练效率	CS:83.5 CV:89.8	–	–
2019	图卷积网络(神经体系结构搜索)[41]	采样和存储效率高	CS:89.4 CV:95.7	Top1:37.1 Top5:60.1	–
2019	图卷积网络(空间残差层、密集连接)[42]	容易与主流时空图卷积方法结合	CS:89.6 CV:95.7	Top1:37.4 Top5:60.4	–
2019	图卷积网络(有向无环图)[43]	识别准确率高	CS:89.9 CV:96.1	Top1:36.9 Top5:59.6	–
2019	共生图卷积网络(行为识别和预测)[44]	增加预测功能，与识别功能相互 促进，提高准确率	CS:90.1 CV:96.4	Top1:37.2 Top5:58.1	–
2020	时空和通道注意的伪图卷积网络[45]	能提取关键帧，但是可能会省略掉部分 关键信息	CS:88.0 CV:93.6	–	–

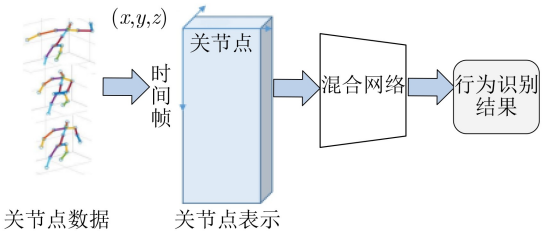

图 6 基于混合网络的关节点行为识别流程图

过增加顶层AGC-LSTM层的时间接受域来增强学习高级特征的能力，从而降低计算成本。Gao等人[52]提出基于双向注意力图卷积网络，利用聚焦和扩散机制从人类关节点数据中学习时空上下文信息，取得了非常好的实验结果，其中在NTU RGB+D数据集上的准确率达到国内外领先水平。Zhang等人[53]将关节的语义(帧索引和关节类型)作为网络输入的一部分与关节的位置和速度一同馈送进语义感知图卷积层和语义感知卷积层，通过实验证明，利用语义信息能够降低模型复杂度和提高行为识别的准确率。利用关节点数据进行行为识别时，骨架关节的复杂时空变化纠缠在一起，Xie等人[54]提出一种时间空间重新校准方案来缓解这种复杂的变化，这是第1次为关节点行为识别开发RNN+CNN网络框架。Weng等人[55]提出一种可变形姿态遍历卷积网络，在执行遍历卷积时通过考虑不同权重的上下文关节来优化每个关节的卷积核大小，对嘈杂的关节更具有鲁棒性，然后将学习的姿势馈送到LSTM共同优化姿势表征和时间序列。主干网络为混合网络的关节点行为识别及代表性工作如表4所示。

3 关节点数据集发展及评估标准

3.1 关节点数据集发展及简述

近年来，深度学习的快速发展使数据驱动学习在行为识别领域取得了较好的成果，大规模的数据集的提出对深度学习的发展有着重大意义。在基于深度学习的关节点行为识别的研究中，相关数据集的发展也同样起着较大的推动作用。在关节点行为识别研究中常用的数据集主要分为两大类，一类是利用Kinect摄像机获取多模态行为识别数据集[9,26,56–69]，另一类是从包含RGB信息的行为识别数据集[68,70,71]中用OpenPose工具箱估计每个关节点的位置[72]。早期的数据集规模较小，相对而言更适用于在深度学习出现之前的手工提取特征方法。同时早期的数据集还有其他的局限性，首先，由于表演者的数量较少和表演者的年龄范围较窄导致行为的内部变化非常有限；其次，行为类别数量较少，通过找到简单的运动模式就可以容易地区分每个行为类别，使分类任务的挑战性降低。为了满足深度学习的需求，大规模数据集相继出现。新加坡南洋理工大学

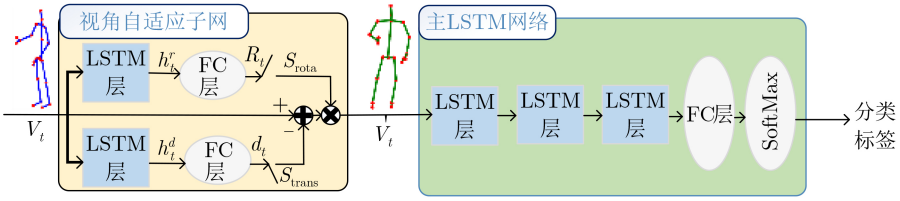

图 7 视图自适应循环神经网络[48]

表 4 主干网络为混合网络的关节点行为识别及代表性工作

年份	技术特点	模型优劣分析	实验结果(%)		
			NTU RGB+D	Kinects	N-UCLA
2018	LSTM+CNN[48]	视图自适应子网减弱了视角变化对识别的影响	CS:88.7 CV:94.3	–	Acc:86.6
2018	CNN+图卷积(多域)[49]	增加了对频率的学习	CS:89.1 CV:94.9	Top1:36.6 Top5:59.1	–
2018	图卷积+LSTM[50]	能同时在空间和时间域上提取行为特征，但模型复杂度较高	CS:84.8 CV:92.4	–	–
2019	图卷积+LSTM(注意力机制)[51]	增加顶层AGC-LSTM层的时间接受域，能够降低计算成本	CS:89.2 CV:95.0	–	–
2019	图卷积+LSTM(双向注意力机制)[52]	非常高的识别准确率	CS:90.3 CV:96.3	Top1:37.3 Top5:60.2	–
2019	图卷积网络(语义)[53]	语义信息能够降低模型复杂度	CS:86.6 CV:93.4	–	Acc:92.5
2018	RNN+CNN[54]	首次采用RNN+CNN的组合提取时空特征，准确率不高	CS:83.0 CV:93.2	–	–
2018	可变形姿势遍历卷积网络+LSTM[55]	对嘈杂关节更具有鲁棒性，但是识别准确率较低	CS:76.8 CV:84.9	–	–

在2016年公开了NTU RGB+D数据集，为国内外进行行为识别研究提供了数据支撑；DeepMind公司在2017年公开了Kinects数据，该数据集从YouTube上收集，以HMDB-51[71]和UCF-101[73]为基准，具有较大的规模和较高的质量。表5列举了常用来做关节点行为识别的多模态数据集，接下来重点介绍在关节点行为识别研究中常用的大规模数据集[26,68,69]。

表 5　关节点行为识别数据集简介

名称	样本数	动作类数	表演者数	视点数	来源	数据形式	年份
Hollywood2[70]	3669	12	--	--	电影	RGB	2009
HMDB[71]	6849	51	--	--	电影	RGB	2011
MSRDailyACtivity3D[56]	320	16	10	1	Kinect v1	RGB/深度/关节点	2011
SBU[57]	300	8	7	3	Kinect v1	RGB/深度/关节点	2012
UT-Kinect[9]	199	10	10	1	Kinect v1	RGB/深度/关节点	2012
3D Action Pairs[58]	360	12	10	1	Kinect v1	RGB/深度/关节点	2013
Florence 3D[59]	215	9	10	1	Kinect v1	RGB/关节点	2013
Multiview 3D Event[60]	3815	8	8	3	Kinect v1	RGB/深度/关节点	2013
Online RGB+D Action[61]	336	7	24	1	Kinect v1	RGB/深度/关节点	2014
N-UCLA[62]	1475	10	10	3	Kinect v1	RGB/深度/关节点	2014
UWA3D [63]	900	30	10	1	Kinect v1	RGB/深度/关节点	2014
UTD-MHAD[64]	861	27	8	1	Kinect v1+传感器	RGB/深度/关节点/惯性传感信号	2015
SYSU 3D[65]	480	12	40	1	Kinect v1	RGB/深度/关节点	2015
UWA 3D Multiview II[66]	1075	30	10	5	Kinect v1	RGB/深度/关节点	2015
M2I[67]	1800	22	22	2	Kinect v1	RGB/深度/关节点	2015
NTU RGB+D[26]	56880	60	40	80	Kinect v2	RGB/深度/关节点/红外信号	2016
Kinects[68]	306245	400	-	-	YouTube	RGB/深度/声音	2017
NTU RGB+D 120[69]	114480	120	106	155	Kinect v2	RGB/深度/关节点/红外信号	2019

NTU RGB+D数据集是由新加坡南洋理工大学制作并整理而成的，于2016年公开。深度传感器的出现使获取物体和人体有效的3D结构的成本大大降低[74]，该数据集是由3个深度摄像机Microsoft Kinect v2在室内拍摄完成的。3个摄像机的水平方向一致，角度分别为–45°, 0°, 45°。包含了25个主要的身体关节的3D位置，如图8所示[26]，关节点位置对照表如表6所示。数据集包含有超过56000个视频样本和400万帧，有40个表演者，60种不同的动作类，涉及日常动作40项(包括饮酒、饮食、阅读等)、交互动作11项(包括拳打脚踢、拥抱等)、与健康相关的动作9项(包括打喷嚏、蹒跚、摔倒等)，数据集示例如图9所示[26]。该数据集有302个样本关节点数据不完整，在进行关节点行为识别时可以忽略。

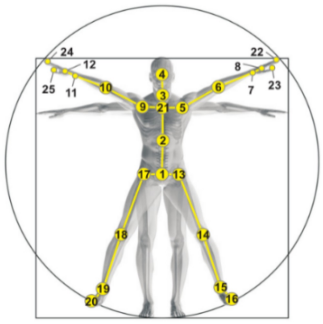

图 8 人体关节点示意图[26]

表 6　关节点位置对照表

序号	对应关节	序号	对应关节	序号	对应关节	序号	对应关节	序号	对应关节
1	脊柱底部	6	左肘	11	右腕	16	左脚	21	脊柱
2	脊柱中间	7	左腕	12	右手	17	右髋	22	左手尖
3	颈	8	左手	13	左髋	18	右膝	23	左手拇指
4	头	9	右肩	14	左膝	19	右踝	24	右手尖
5	左肩	10	右肘	15	左踝	20	右脚	25	右手拇指

Kinects数据集取自YouTube视频，每段动作剪辑约10 s，包含400个动作类，每个动作类由400～1150个视频剪辑。动作涵盖范围较广，包含人与物的交互、人与人的交互、单人动作。利用公开的Openpose工具箱能够在Kinects数据集提取18个关节点位置(X,Y,C)，其(X,Y)为关节点的2维位置坐标，C是位置坐标的置信度，关节框架被记录为18个元组的数组，图10为Openpose工具箱提取关节点示意图[72]。

NTU RGB+D 120数据集在NTU RGB+D数据集的基础上扩充到了120个动作，动作的种类未发生变化，每个动作类包含的动作个数均有增加，日常动作增加到了82个(包括吃、写、坐下、移动物体等)，与健康有关的动作增加到了12个(包括吹鼻子、呕吐、蹒跚、跌倒等)，交互的动作增加到了26个(包括握手、推、打、拥抱等)。与NTU RGB+D数据集相比，该数据集行为识别的难度有所增加。

3.2 关节点数据集评估标准

常见行为识别准确率的评估标准为Top1和Top5。模型预测某个行为类别的准确率时，如M2I数据集包含行为类别有22个，模型会给出22个按概率从高到低的类别排名。其中Top1的准确率为排名第1的类别与实际结果相符的准确率，Top5的准确率为排名前5类别中包含实际结果的准确率。一般一种模型在一个数据集上实验结果的准确率Acc即为Top1的准确率。NTU RGB+D和NTU RGB+D 120数据集经常出现CS和CV两种测试模式，其中CS为跨表演者测试中Top1的准确率，CV为跨视角测试中Top1的准确率。Kinects数据集较为特殊，对其而言Top5比Top1更有说服力，因为该数据集中一段视频可能包含多个动作但是标签仅标注一个动作，因此在表1—表4中Kinects数据集上的实验结果同时包含Top1和Top5。

(a) 握手动作 (b) 合掌动作

图 9 NTU RGB+D数据集示例[26]

图 10 Openpose提取关节点示意图[72]

4 总结与展望

本文通过对基于深度学习的关节点行为识别进行总结和分析，得出以下结论：

(1) 关节点数据一般有3种表示方式：伪图像、向量序列和拓扑图。卷积神经网络适合处理伪图像，循环神经网络适合处理向量序列，图卷积网络适合处理拓扑图。从表1—表4可知，在NTU RGB+D数据集上仅包含图卷积的方法比仅比包含卷积神经网络的方法平均准确率高约5%，比仅包含循环神经网络的方法高约10%，证明了图卷积在关节点行为识别方面的优越性，这是因为人体关节点所构成的骨架实质上就是图结构，但是通常包含图卷积的网络也更复杂。仅包含循环神经网络方法的识别准确率相对最低，因为行为动作在空间域上的变化幅度要大于在时间域上的，而循环神经网络更适合处理时序性问题。基于混合网络的关节点行为识别方法通常具有两种或多种主干网络的优势其准确率也较高，但同时网络的复杂度也较高。

(2) 目前关节点行为识别方法在标准数据集上的准确率有大幅度提高，以NTU RGB+D数据集为例，已经从CS和CV的准确率分别为62.9%和70.3%[26]提升到了90.3%和96.3%[52]。但实际应用场景中可能会出现关节点的部分数据缺失或需要较高的训练和测试速度以实现实时性，有些学者针对这些问题提出了解决方法，比如缺失关节点的激活[40]或者构建轻量级的网络[24]，但目前准确率都一般。也有些研究是针对特殊的应用场景，比如刘庭煜等人[19]针对车间工人行为识别，虽然在特定的场景中准确率较高，但适用性较差。

综合当前基于深度学习的关节点行为识别方法的研究现状，对今后的研究做出如下展望：

(1) 随着5G时代的到来，数据信息的主要载体已经从PC转换到移动端，这有利用将关节点行为识别应用于移动端。但是目前利用关节点进行行为识别的网络模型均较为复杂，其中以循环神经网络和图卷积网络最为明显，难以在实际应用中推广，因此期待未来能够提出更加轻量级并且准确度较高的网络。

(2) 关节点行为识别多应用于无人驾驶、机器人以及医疗监控等领域，行为识别系统在行为动作发生之后对行为进行识别。但是在某些应用场景中人们更希望能够进行行为预测，比如当无人驾驶系统预测到一个人有闯红灯的行为时可以及时调整驾驶轨迹。Li等人[44]就利用关节点的行为预测进行了

深入的研究，但是准确率有待提高，这也是未来的研究方向之一。

(3) 目前关节点行为识别的训练数据多是剪辑好的视频帧，无需进行动作检测，但是在实际应用中，能够识别行为发生的时间段是有必要的，因此将关节点的动作检测与行为识别相结合也有较高的研究价值。

(4) 深度学习需要大量的样本进行训练，但对数据集进行准确有效的标注是需要耗费大量人力物力的。无监督学习可以利用无标签的数据进行训练，这将解决数据集标注所面临的问题，具有较大的研究价值。

(5) 虽然很多关节点行为识别方法在标准数据集上识别的准确率很高，但是这些方法都是针对无遮挡的情况进行的，在实际的应用场景中可能会出现部分关节点被遮挡的情况，现在的大部分方法在这种情况下的识别效果并不好，因此提高在有遮挡情况下的识别准确率有利于关节点行为识别与实际应用的结合。

参 考 文 献

[1] 吴培良, 杨霄, 毛秉毅, 等. 一种视角无关的时空关联深度视频行为识别方法[J]. 电子与信息学报, 2019, 41(4): 904–910. doi: 10.11999/JEIT180477.
WU Peiliang, YANG Xiao, MAO Bingyi, *et al*. A perspective-independent method for behavior recognition in depth video via temporal-spatial correlating[J]. *Journal of Electronics & Information Technology*, 2019, 41(4): 904–910. doi: 10.11999/JEIT180477.

[2] 朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6): 848–857. doi: 10.16383/j.aas.2016.c150710.
ZHU Yu, ZHAO Jiangkun, WANG Yining, *et al*. A review of human action recognition based on deep learning[J]. *Acta Automatica Sinica*, 2016, 42(6): 848–857. doi: 10.16383/j.aas.2016.c150710.

[3] 罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 169–180. doi: 10.11959/j.issn.1000-436x.2018107.
LUO Huilan, WANG Chanjuan, and LU Fei. Survey of video behavior recognition[J]. *Journal on Communications*, 2018, 39(6): 169–180. doi: 10.11959/j.issn.1000-436x.2018107.

[4] 张会珍, 刘云麟, 任伟建, 等. 人体行为识别特征提取方法综述[J]. 吉林大学学报: 信息科学版, 2020, 38(3): 360–370.
ZHANG Huizhen, LIU Yunlin, REN Weijian, *et al*. Human behavior recognition feature extraction method: A survey[J]. *Journal of Jilin University: Information Science Edition*, 2020, 38(3): 360–370.

[5] ZHU Fan, SHAO Ling, XIE Jin, *et al*. From handcrafted to learned representations for human action recognition: A survey[J]. *Image and Vision Computing*, 2016, 55(2): 42–52. doi: 10.1016/j.imavis.2016.06.007.

[6] ZHANG Zhengyou. Microsoft kinect sensor and its effect[J]. *IEEE Multimedia*, 2012, 19(2): 4–10. doi: 10.1109/MMUL.2012.24.

[7] YAN Yichao, XU Jingwei, NI Bingbing, *et al*. Skeleton-aided articulated motion generation[C]. The 25th ACM International Conference on Multimedia, Mountain View, USA, 2017: 199–207. doi: 10.1145/3123266.3123277.

[8] HAN Fei, REILY B, HOFF W, *et al*. Space-time representation of people based on 3D skeletal data: A review[J]. *Computer Vision and Image Understanding*, 2017, 158: 85–105. doi: 10.1016/j.cviu.2017.01.011.

[9] XIA Lu, CHEN C C, and AGGARWAL J K. View invariant human action recognition using histograms of 3D joints[C]. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, USA, 2012: 20–27.

[10] WENG Junwu, WENG Chaoqun, and YUAN Junsong. Spatio-temporal Naive-Bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Tianjin, China, 2017: 4171–4180.

[11] LI Bo, DAI Yuchao, CHENG Xuelian, *et al*. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN[C]. 2017 IEEE International Conference on Multimedia & Expo Workshops, Hong Kong, China, 2017: 4171–4180. doi: 10.1109/ICMEW.2017.8026282.

[12] KIM T S and REITER A. Interpretable 3D human action analysis with temporal convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 1623–1631. dio: 10.1109/CVPRW. 2017.207.

[13] LI Chao, ZHONG Qiaoyong, XIE Di, *et al*. Skeleton-based action recognition with convolutional neural networks[C]. 2017 IEEE International Conference on Multimedia & Expo Workshops, Hong Kong, China, 2017: 597–600. doi: 10.1109/ICMEW.2017.8026285.

[14] LIU Mengyuan, LIU Hong, and CHEN Chen. Enhanced skeleton visualization for view invariant human action recognition[J]. *Pattern Recognition*, 2017, 68: 346–362. doi: 10.1016/j.patcog.2017.02.030.

[15] KE Qiuhong, AN Senjian, BENNAMOUN M, *et al*. SkeletonNet: Mining deep part features for 3-D action recognition[J]. *IEEE Signal Processing Letters*, 2017, 24(6): 731–735. doi: 10.1109/LSP.2017.2690339.

[16] KE Qiuhong, BENNAMOUN M, AN Senjian, *et al*. A new representation of skeleton sequences for 3D action

recognition[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 3288–3297.

[17] LE T M, INOUE N, and SHINODA K. A fine-to-coarse convolutional neural network for 3D human action recognition[J]. arXiv preprint arXiv: 1805.11790, 2018.

[18] LI Chao, ZHONG Qiaoyong, XIE Di, *et al*. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[J]. arXiv preprint arXiv: 1804.06055, 2018.

[19] 刘庭煜, 陆增, 孙毅锋, 等. 基于三维深度卷积神经网络的车间生产行为识别[J]. 计算机集成制造系统, 2020, 26(8): 2143–2156.
LIU Tingyu, LU Zeng, SUN Yifeng, *et al*. Working activity recognition approach based on 3D deep convolutional neural network[J]. *Computer Integrated Manufacturing Systems*, 2020, 26(8): 2143–2156.

[20] 姬晓飞, 秦琳琳, 王扬扬. 基于RGB和关节点数据融合模型的双人交互行为识别[J]. 计算机应用, 2019, 39(11): 3349–3354. doi: 772/j.issn.1001-9081.2019040633.
JI Xiaofei, QIN Linlin, and WANG Yangyang. Human interaction recognition based on RGB and skeleton data fusion model[J]. *Journal of Computer Applications*, 2019, 39(11): 3349–3354. doi: 772/j.issn.1001-9081.2019040633.

[21] YAN An, WANG Yali, LI Zhifeng, *et al*. PA3D: Pose-action 3D machine for video recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 7922–7931. doi: 10.1109/CVPR.2019.00811.

[22] CAETANO C, BRÉMOND F, and SCHWARTZ W R. Skeleton image representation for 3D action recognition based on tree structure and reference joints[C]. The 32nd SIBGRAPI Conference on Graphics, Patterns and Images, Rio de Janeiro, Brazil, 2019: 16–23.

[23] CAETANO C, SENA J, BRÉMOND F, *et al*. SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition[C]. The 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, Taipei, China, 2019: 1–8. doi: 10.1109/AVSS.2019.8909840.

[24] LI Yanshan, XIA Rongjie, LIU Xing, *et al*. Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition[C]. 2019 IEEE International Conference on Multimedia and Expo, Shanghai, China, 2019: 1066–1071. doi: 10.1109/ICME.2019.00187.

[25] YANG Fan, WU Yang, SAKTI S, *et al*. Make skeleton-based action recognition model smaller, faster and better[C]. The ACM Multimedia Asia, Beijing, China, 2019: 1–6. doi: 10.1145/3338533.3366569.

[26] SHAHROUDY A, LIU Jun, NG T T, *et al*. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1010–1019. doi: 10.1109/CVPR.2016.115.

[27] LIU Jun, SHAHROUDY A, XU Dong, *et al*. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]. The European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 816–833. doi: 10.1007/978-3-319-46487-9_50.

[28] LIU Jun, SHAHROUDY A, XU Dong, *et al*. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 40(12): 3007–3021. doi: 10.1109/TPAMI.2017.2771306.

[29] LIU Jun, WANG Gang, HU Ping, *et al*. Global context-aware attention LSTM networks for 3D action recognition[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1647–1656. doi: 10.1109/CVPR.2017.391.

[30] LIU Jun, WANG Gang, DUAN Lingyun, *et al*. Skeleton-based human action recognition with global context-aware attention LSTM networks[J]. *IEEE Transactions on Image Processing*, 2018, 27(4): 1586–1599.

[31] ZHENG Wu, LI Lin, ZHANG Zhaoxiang, *et al*. Relational network for skeleton-based action recognition[C]. 2019 IEEE International Conference on Multimedia and Expo, Shanghai, China, 2019: 826–831.

[32] LI Shuai, LI Wanqing, COOK C, *et al*. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5457–5466. doi: 10.1109/CVPR.2018.00572.

[33] 王佳铖, 鲍劲松, 刘天元, 等. 基于工件注意力的车间作业行为在线识别方法[J/OL]. 计算机集成制造系统, 2020: 1–13. http://kns.cnki.net/kcms/detail/11.5946.TP.20200623.1501.034.html.
WANG Jiacheng, BAO Jinsong, LIU Tianyuan, *et al*. Online method for worker operation recognition based on the attention of workpiece[J/OL]. *Computer Integrated Manufacturing Systems*, 2020: 1–13. http://kns.cnki.net/kcms/detail/11.5946.TP.20200623.1501.034.html.

[34] YAN Sijie, XIONG Yuanjun, and LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[J]. arXiv preprint arXiv: 1801.07455, 2018.

[35] SHI Lei, ZHANG Yifan, CHENG Jia, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 12026–12035.

[36] LI M, CHEN Siheng, CHEN Xu, *et al*. Actional-structural graph convolutional networks for skeleton-based action recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 3595–3603.

[37] GAO Xiang, HU Wei, TANG Jiaxiang, *et al*. Optimized skeleton-based action recognition via sparsified graph regression[C]. The 27th ACM International Conference on Multimedia, New York, USA, 2019: 601–610.

[38] LI Chaolong, CUI Zhen, ZHENG Wenming, *et al*. Spatio-temporal graph convolution for skeleton based action recognition[C]. The 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 247–254.

[39] TANG Yansong, TIAN Yi, LU Jiwen, *et al*. Deep progressive reinforcement learning for skeleton-based action recognition[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5323–5332.

[40] SONG Yifan, ZHANG Zhang, and WANG Liang. Richly activated graph convolutional network for action recognition with incomplete skeletons[C]. 2019 IEEE International Conference on Image Processing, Taipei, China, 2019: 1–5. doi: 10.1109/ICIP.2019.8802917.

[41] PENG Wei, HONG Xiaopeng, CHEN Haoyu, *et al*. Learning graph convolutional network for skeleton-based human action recognition by neural searching[J]. arXiv preprint arXiv: 1911.04131, 2019.

[42] WU Cong, WU Xiaojun, and KITTLER J. Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition[C]. 2019 IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea, 2019: 1–5.

[43] SHI Lei, ZHANG Yifan, CHENG Jian, *et al*. Skeleton-based action recognition with directed graph neural networks[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 7912–7921.

[44] LI Maosen, CHEN Siheng, CHEN Xu, *et al*. Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction[J]. arXiv preprint arXiv: 1910.02212, 2019.

[45] YANG Hongye, GU Yuzhang, ZHU Jianchao, *et al*. PGCN-TCA: Pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition[J]. *IEEE Access*, 2020, 8: 10040–10047. doi: 10.1109/ACCESS.2020.2964115.

[46] WU Felix, ZHANG Tianyi, DE SOUZA JR A H, *et al*. Simplifying graph convolutional networks[J]. arXiv preprint arXiv: 1902.07153, 2019.

[47] CHEN Jie, MA Tengfei, and XIAO Cao. FastGCN: Fast learning with graph convolutional networks via importance sampling[J]. arXiv preprint arXiv: 1801.10247, 2018.

[48] ZHANG Pengfei, LAN Cuiling, XING Junliang, *et al*. View adaptive neural networks for high performance skeleton-based human action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(8): 1963–1978. doi: 10.1109/TPAMI.2019.2896631.

[49] HU Guyue, CUI Bo, and YU Shan. Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention[C]. 2019 IEEE International Conference on Multimedia and Expo, Shanghai, China, 2019: 1216–1221.

[50] SI Chenyang, JING Ya, WANG Wei, *et al*. Skeleton-based action recognition with spatial reasoning and temporal stack learning[C]. The European Conference on Computer Vision, Munich, Germany, 2018: 103–118.

[51] SI Chenyang, CHEN Wentao, WANG Wei, *et al*. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 1227–1236. doi: 10.1109/CVPR.2019.00132.

[52] GAO Jialin, HE Tong, ZHOU Xi, *et al*. Focusing and diffusion: Bidirectional attentive graph convolutional networks for skeleton-based action recognition[J]. arXiv preprint arXiv: 1912.11521, 2019.

[53] ZHANG Pengfei, LAN Cuiling, ZENG Wenjun, *et al*. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020. doi: 10.1109/CVPR42600.2020.00119.

[54] XIE Chunyu, LI Ce, ZHANG Baochang, *et al*. Memory attention networks for skeleton-based action recognition[C]. The 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018.

[55] WENG Junwu, LIU Mengyuan, JIANG Xudong, *et al*. Deformable pose traversal convolution for 3D action and gesture recognition[C]. The European Conference on Computer Vision, Munich, Germany, 2018: 768–775. doi: 10.1007/978-3-030-01234-2_9.

[56] WANG Jiang, LIU Zicheng, WU Ying, *et al*. Mining actionlet ensemble for action recognition with depth cameras[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 1290–1297. doi: 10.1109/CVPR.2012.6247813.

[57] YUN K, HONORIO J, CHATTOPADHYAY D, *et al*. Two-person interaction detection using body-pose features and multiple instance learning[C]. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, USA, 2012: 28–35. doi:

10.1109/CVPRW.2012.6239234.

[58] OREIFEJ O and LIU Zicheng. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences[C]. 2013 IEEE conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 716–723. doi: 10.1109/CVPR.2013.98.

[59] SEIDENARI L, VARANO V, BERRETTI S, *et al*. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses[C]. 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, USA, 2013: 479–485.

[60] WEI Ping, ZHAO Yibiao, ZHENG Nanning, *et al*. Modeling 4D human-object interactions for event and object recognition[C]. 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 2013: 3272–3279.

[61] YU Gang, LIU Zicheng, and YUAN Junsong. Discriminative orderlet mining for real-time recognition of human-object interaction[C]. The Asian Conference on Computer Vision, Singapore, 2014: 50–65. doi: 10.1007/978-3-319-16814-2_4.

[62] WANG Jiang, NIE Xiaohan, XIA Yin, *et al*. Cross-view action modeling, learning, and recognition[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 2649–2656. doi: 10.1109/CVPR.2014.339.

[63] RAHMANI H, MAHMOOD A, HUYNH D Q, *et al*. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition[C]. The European Conference on Computer Vision, Zurich, Switzerland, 2014: 742–757. doi: 10.1007/978-3-319-10605-2_48.

[64] CHEN Chen, JAFARI R, and KEHTARNAVAZ N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor[C]. 2015 IEEE International Conference on Image Processing, Quebec, Canada, 2015: 168–172.

[65] HU Jianfang, ZHENG Weishi, LAI Jianhuang, *et al*. Jointly learning heterogeneous features for RGB-D activity recognition[C]. 2015 IEEE conference on Computer Vision and Pattern Recognition, Boston, America, 2015: 5344–5352.

[66] RAHMANI H, MAHMOOD A, HUYNH D, *et al*. Histogram of oriented principal components for cross-view action recognition[J]. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(12): 2430–2443. doi: 10.1109/TPAMI.2016.2533389.

[67] XU Ning, LIU Anan, NIE Weizhi, *et al*. Multi-modal & multi-view & interactive benchmark dataset for human action recognition[C]. The 23rd ACM International Conference on Multimedia, Brisbane, Australia, 2015: 1195–1198.

[68] KAY W, CARREIRA J, SIMONYAN K, *et al*. The kinetics human action video dataset[J]. arXiv preprint arXiv: 1705.06950, 2017.

[69] LIU Jun, SHAHROUDY A, PEREZ M, *et al*. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(10): 2684–2701. doi: 10.1109/TPAMI.2019.2916873.

[70] MARSZALEK M, LAPTEV I, and SCHMID C. Actions in context[C]. 2019 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 2391–2396. doi: 10.1109/CVPR.2009.5206557.

[71] KUEHNE H, JUANG H, GARROTE E, *et al*. HMDB: A large video database for human motion recognition[C]. 2011 International Conference on Computer Vision, Barcelona, Spain, 2011: 2556-2563. doi:10.1007/978-3-642-33374-3_41 .

[72] CAO Zhe, SIMON T, WEI S E, *et al*. Realtime multi-person 2D pose estimation using part affinity fields[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017: 7291–7299. doi: 10.1109/CVPR.2017.143.

[73] SOOMRO K, ZAMIR A R, and SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv: 1212.0402, 2012.

[74] HAN Jungong, SHAO Ling, XU Dong, *et al*. Enhanced computer vision with microsoft kinect sensor: A review[J]. *IEEE Transactions on Cybernetics*, 2013, 43(5): 1318–1334. doi: 10.1109/TCYB.2013.2265378.

刘 云：男，1962年生，教授，研究方向为计算机视觉.
薛盼盼：女，1995年生，硕士生，研究方向为计算机视觉.
李 辉：男，1984年生，副教授，研究方向为计算机视觉.
王传旭：男，1968年生，教授，研究方向为计算机视觉.

责任编辑：余 蓉